Technological Improvements or Climate Change? Bayesian Modeling of Time-Varying Conformance to Benford's Law

Technological Improvements or Climate Change? Bayesian Modeling of Time-Varying Conformance to Benford's Law

Junho Lee, Miguel de Carvalho*,

School of Mathematics, University of Edinburgh, Edinburgh, UK

* Miguel.deCarvalho@ed.ac.uk

Abstract

We develop a Bayesian time-varying model that tracks periods at which conformance to Benford's Law is lower. Our methods are motivated by recent attempts to assess how the quality and homogeneity of large datasets may change over time by using the First-Digit Rule. We resort to a smooth multinomial logistic model which captures the dynamics governing the proportion of first digits, and apply the proposed model to global tropical cyclone tracks over the past two centuries. Our findings indicate that cumulative technological improvements may have only had a moderate influence on the homogeneity of the dataset, and hint that recent heterogeneity could be due to other drivers.

Introduction

Benford's Law is an empirical observation on the distribution of first digits of numerical data discovered by [1] and [2]. The law states that, in many situations of applied interest, the frequency of the first digit of numbers follows a logarithmically decreasing distribution—even though it is generally believed that the probability of occurrence of each number is equally likely. The probability that the first non-zero digit begins with a number d follows a logarithmic distribution given by

$$\mathsf{p}_d = \mathsf{P}(D = d) = \log_{10}\left(1 + \frac{1}{d}\right), \quad d = 1, \ldots, 9, \tag{1}$$

where D is the first significant digit of a random variable. The probability of the significant leading digit equal to 1, for example, is calculated as approximately 0.301, and then the probability of the leading digit equals d gets smaller as d increase, up to where the probability of the leading digit 9 equals to only 0.046. A wide variety of datasets, especially a collection of datasets, have been reported to conform to Benford's Law. A statistical foundation of its universality was presented by [3]. Since the peculiar law of first digits uncovered, a battery of studies showed that large classes of quantities in different disciplines from both natural phenomena and social activities are expected to follow the First-Digit Rule, and therefore it can be used for detecting structural changes or irregularities from various applications [4,5].

This paper devises a Bayesian time-varying model that tracks periods at which conformance to Benford's Law is lower. Our methods are motivated by recent attempts to assess how the quality and homogeneity of large datasets may change over time by using the First-Digit Rule (e.g. [6–8]). As we show in the numerical studies in the Supplementary Materials (S1 File), the empirical-based approach by [8] suffers often from bias (cf Fig 4, S1 File)—thus questioning some of their key empirical findings. Our Bayesian smooth multinomial logistic model is however accurate (cf numerical studies in S1 File), and it is tailored—by construction—for capturing the dynamics governing the proportion of first digits. We apply the proposed model to global tropical cyclone tracks over the past two centuries, and compare our empirical findings with those of [8]. An application of our model indicates that cumulative technological improvements may have only had a moderate influence on the homogeneity of the dataset. Indeed, although technological improvements are cumulative we find that the most recent heterogeneity levels actually tend to be higher than the ones from 1842 to 1890 (cf Fig 4, below); this finding seems to be in contradiction with [8] (cf Fig 5 in their paper), possibly due to the above-mentioned bias issue. Finally, while we center the article on the tropical cyclone application, our Bayesian time-varying approach has the potential to be employed on other contexts where the target is on learning about the dynamics governing conformance to Benford's Law—including fraud analysis.

The paper is organized as follows. We first introduce our motivating global tropical cyclone data and provide preliminary statistics on their conformance to Benford's Law. The next section describes our proposed Bayesian multinomial logistic smoothing model

along with details on prior specification and on inference. The homogeneity of cyclone data is then analyzed by inspecting dynamics of the first-digit distribution. Lastly, we discuss data homogeneity and other issues based on the results. For the convenience of exposition, specific details surrounding numerical experiments on the model and relevant code in R [9] are left to the S1 File.

Materials and Methods

Global Tropical Cyclones (GTC) Dataset

Fig 1. Map of the global tropical cyclones tracks from International Best Track Archive for Climate Stewardship (IBTrACS).
(Top) paths from 1842 to 1960; (Bottom) paths from 1961 to 2017;(Left) Map projection on longitude 10°E ∼ 170°W; (Right) Map projection on Longitude 170°W ∼ 10°E.

The GTC dataset provides information on the distribution, frequency, and intensity of tropical cyclones worldwide, which is collected as a project of International Best Track Archive for Climate Stewardship (IBTrACS). The dataset includes a register of tropical cyclones since 1842, and is available from the website of IBTrACS (https://www.ncdc.noaa.gov/ibtracs). It has multiple observed records of each cyclone such as geographical location, temperature, and wind speed. As of May of 2018, a total of 348,703 traveled locations are recorded with the corresponding climatic information. Fig 1 presents the traveled path of each cyclone in the dataset over the entire period.

Fig 2. Descriptive statistics for GTC data.
Above: Frequency of tropical cyclones from 1842–2016. Below: Empirical first-digit distribution of the traveled distance (in meters) per cyclone is represented (red) along with the corresponding probability mass for Benford's distribution (blue).

Apart from the intrinsic heterogeneity of tropical cyclones, there has been a debate on the quality of early records in the dataset for assessing the influence of climate change on the occurrence of tropical cyclones [10,11].

We retrieve observed location records of each cyclone from 1842 to 2016 in the GTC dataset and then trace a geometric path by connecting points which each cyclone traveled. We measure distance per cyclone in meters along the path using the `geosphere` package [12] from the R programming language. Except for the small cyclones with a single geographical location (latitude/longitude), we obtain 12,741

observations of traveled distances in total; Fig 2 depicts the frequency of tropical cyclones over the period under analysis. This allows us to analyze dynamics of the first-digit proportion throughout the period under analysis. Before specifying our statistical model, we first test the overall validity of Benford's Law in the GTC dataset. Fig 2 shows the proportion of first digits in the GTC dataset against Benford's Law. The first-digit proportions in the pooled GTC data resemble the probability mass from Benford's Law. In Fig 2, digit 1 and digit 6 to digit 9 exhibits higher proportion than the probability from Equation (1) among all digits, whereas digits 2 to 4 present lower than the counterpart values. We discuss the variation of each proportion in detail with our time-varying model.

Modeling Time-varying Conformance to Benford's Law

Model Specification

We construct a smooth multinomial model which will capture the time-varying proportions in the leading digits and compare the variation with Benford's distribution. Our GTC dataset is composed of two records for each cyclone: the first digit of traveled distance and the year a cyclone was first observed. Let N_t be the number of cyclones occurring in year t, and let $\boldsymbol{n}_t = (n_{1,t}, \ldots, n_{9,t})$ with $n_{d,t}$ denoting the frequency of cyclones whose first digit of traveled distance equals to d, during year t. Below, $\boldsymbol{n}_t$ is assumed to follow a multinomial distribution with parameter

$$(N_t, p_{1,t}, \ldots, p_{9,t}), \tag{2}$$

where the $p_{d,t}$'s obey $\sum_{d=1}^{9} p_{d,t} = 1$, for all t.

Our primary interest is in the probability $\boldsymbol{p}_t = (p_{1,t}, \ldots, p_{9,t})$, that is the probability of occurrence of each digit at year t; we will refer to $\boldsymbol{p}_t$ as the *first-digit probability*. More precisely, our target below will be on learning from the data about the dynamics governing the first-digit probability, $\boldsymbol{p}_t$, and on contrasting it with Benford's Law, p_d, in Equation (1).

Since our data is composed of frequencies of nine digits together with time t, it is natural to relate the first-digit probability to the time predictor via a generalized linear

model [13]. We consider a multinomial logistic model where elements of $\boldsymbol{p}_t$ are connected to a vector of time predictor $\boldsymbol{\eta}_t = (\eta_{1,t}, \ldots, \eta_{8,t})$ by

$$p_{d,t} = \frac{\exp(\eta_{d,t})}{1 + \sum_{i=1}^{8} \exp(\eta_{i,t})}, \quad d = 1, \ldots, 8, \tag{3}$$

and $p_{9,t}$ is inferred from $\sum_{d=1}^{9} p_{d,t} = 1$. Time-varying conformance to Benford's Law will then be assessed by contrasting $p_{d,t}$ as in (3) against the benchmark p_d, from Equation (1).

To trace the dynamics governing $\boldsymbol{p}_t$, we employ degree 3 B-spline basis [14], also known as cubic splines, which produce a smooth curve for each element of $\boldsymbol{\eta}_t$; cubic splines are the standard choice in the literature as they are twice continuously differentiable and thus allow for a reasonable amount of smoothness [15]. We assume that the B–spline basis functions have $K + 1$ equally spaced knots, $t_{\min} = t_0 < t_1 < \cdots < t_{K-1} < t_K = t_{\max}$ over the entire observation period, and thus the smooth curve $\eta_{d,t}$ can be expressed by the following linear combination of B–splines,

$$\eta_{d,t} = \sum_{k=1}^{K+3} \beta_{d,k} B_k(t), \quad d = 1, \ldots, 8. \tag{4}$$

Here the $\beta_{d,k}$'s are regression coefficients of B-splines predictors for digit d, and $B_k(t)$ is a set of B–splines basis functions of degree 3.

To assess overall conformance over nine digits with the First-Digit Rule in each year, we use the smooth sum of squared deviations (SSD) of each digit as a summary statistic. The smooth SSD is computed by a sum of squares of the individual discrepancies between leading digits, i.e.

$$\mathrm{SSD}(t) = \sum_{d=1}^{9} (p_{d,t} - \mathsf{p}_d)^2, \tag{5}$$

where $p_{d,t}$ and p_d are respectively the first-digit probability from (3), and the probability from Benford's Law from (1). The smooth SSD will be exactly zero when the first-digit probability happens to equal to Benford's first-digit distribution.

To sum up, the goal of the model is on tracking the dynamics governing the first-digit probability over time, conformance to the benchmark will be assessed via the

smooth SSD as in (5), and we next concentrate on discussing how the Bayesian paradigm can be used to learn about $\boldsymbol{p}_t$ from the data.

Bayesian Inference

We follow a Bayesian version of the penalised spline approach [16,17] so as to learn about the first-digit probability $\boldsymbol{p}_t$. We assign a first-order random walk prior to the regression coefficients $\boldsymbol{\beta}_d = (\beta_{1,d}, \ldots, \beta_{K+3,d})^{\mathrm{T}}$, which relate an independent and identical Gaussian error ε_d with mean zero and variance τ_d^2, that is,

$$\beta_{d,k} = \beta_{d,k-1} + \varepsilon_d, \quad \varepsilon_d \sim N(0, \tau_d^2), \quad k = 2, \ldots, K+3; \tag{6}$$

a flat (uniform) prior is set for the initial coefficient $\beta_{d,1}$. The first order random walk prior can be represented in a matrix form, $\boldsymbol{F}\boldsymbol{\beta}_d = \boldsymbol{\varepsilon}_d$, where $\boldsymbol{\varepsilon}_d$ is a $(K+2)$-vector of ε_d's and $\boldsymbol{F}$ is a difference matrix of dimension $(K+2, K+3)$. The $\boldsymbol{F}$ has a diagonal of 1's $(i = j)$, -1's for the next elements to the diagonal $(i = j+1)$, and zero otherwise for the (i, j)th element with $i \in \{1, \ldots, K+2\}$ and $j \in \{1, \ldots, K+3\}$.

The variance τ_d^2 controls amount of smoothness of $\eta_{d,t}$—and hence that of $p_{d,t}$—with a lower τ_d^2 indicating that variability of the next regression coefficient is restricted around the value of the previous coefficient. Accordingly, the conditional probability of the regression coefficients $\boldsymbol{\beta}_d$ given τ_d^2 is given by

$$\pi(\boldsymbol{\beta}_d \mid \tau_d^2) \propto \exp\left(-\frac{1}{2\tau_d^2}\boldsymbol{\beta}_d^{\mathrm{T}}\boldsymbol{K}\boldsymbol{\beta}_d\right), \tag{7}$$

where $\boldsymbol{K}$ is a penalty matrix, $\boldsymbol{K} = \boldsymbol{F}^{\mathrm{T}}\boldsymbol{F}$ obtained from the random walk prior in Equation (6). The precision parameters τ_d^2's are estimated along with the regression coefficients in the model by assigning an additional prior. We place a diffuse inverse gamma prior $\tau_d^2 \sim \mathrm{IG}(a_0, b_0)$ with two constants a_0 and b_0 and then apply a uniform prior for performing a sensitivity analysis. To ease notation, in what follows we let $\boldsymbol{\beta}$ and $\boldsymbol{\tau}$ stand for the set $\{\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_8\}$ and $\{\boldsymbol{\tau}_1^2, \ldots, \boldsymbol{\tau}_8^2\}$ respectively.

The likelihood of observing $\boldsymbol{n} = \{\boldsymbol{n}_1, \ldots, \boldsymbol{n}_T\}$ is given by the product of multinomial

probabilities, that is,

$$L(\boldsymbol{\beta}) = f(\boldsymbol{n}_1, \ldots, \boldsymbol{n}_T \mid \boldsymbol{p}_1, \ldots, \boldsymbol{p}_T) \propto \prod_{t=1}^{T} \prod_{d=1}^{9} \{p_{d,t}(\boldsymbol{\beta})\}^{n_{d,t}}, \tag{8}$$

where $p_{d,t}(\boldsymbol{\beta})$ and $n_{d,t}$ are respectively the probability and the realized frequency of digit d in year t; note that $p_{d,t}$ is connected to the regression coefficients $\boldsymbol{\beta}$ via the link function in (3) and the linear predictors $\eta_{d,t}$ in (4). The model is summarized in Box 1.

Bayesian multinomial logistic smoothing model

$$\begin{aligned}
&\text{(Likelihood)} && (n_{1,t}, \ldots, n_{9,t}) \sim \text{Mult}(N_t, p_{1,t}, \ldots, p_{9,t}),\\
&\text{(Model Specification)} && p_{d,t} = \frac{\exp(\eta_{d,t})}{1 + \sum_{d=1}^{8} \exp(\eta_{d,t})}, \quad p_{9,t} = \frac{1}{1 + \sum_{d=1}^{8} \exp(\eta_{d,t})},\\
& && \eta_{d,t} = \sum_{k=1}^{K+3} \beta_{d,k} B_{d,k}(t),\\
&\text{(Random Walk Prior)} && \beta_{1,d} \sim U(c_0, d_0), \quad \beta_{k+1,d} = \beta_{k,d} + \varepsilon_d, \quad \varepsilon_d \sim N(0, \tau_d^2),\\
&\text{(Hyper-Prior)} && \tau_d^2 \sim \text{IG}(a_0, b_0).
\end{aligned}$$

Box 1. Summary description of the fitted Bayesian smoothing model.

Bayesian inference is based on the joint posterior distribution given by

$$\mathsf{p}(\boldsymbol{\beta}, \boldsymbol{\tau}^2 \mid \boldsymbol{n}) \propto L(\boldsymbol{\beta})\, \pi(\boldsymbol{\beta} \mid \boldsymbol{\tau}^2)\, \pi(\boldsymbol{\tau}^2), \tag{9}$$

where $\pi(\boldsymbol{\tau}^2) = \prod_{d=1}^{8} \pi(\tau_d)$, with $\pi(\tau_d)$ denoting the density of an inverse gamma distribution with parameters (a_0, b_0), and $\pi(\boldsymbol{\beta} \mid \boldsymbol{\tau}^2) = \prod_{d=1}^{8} \pi(\boldsymbol{\beta}_d \mid \tau_d^2)$ with $\pi(\boldsymbol{\beta}_d \mid \tau_d^2)$ as in (7). We calculate a full conditional distribution for the regression coefficients $\boldsymbol{\beta}$ and $\boldsymbol{\tau}^2$ from Equation (9),

$$\mathsf{p}(\boldsymbol{\beta} \mid \boldsymbol{n}, \boldsymbol{\tau}^2) \propto L(\boldsymbol{\beta})\, \pi(\boldsymbol{\beta} \mid \boldsymbol{\tau}^2), \quad \mathsf{p}(\boldsymbol{\tau}^2 \mid \boldsymbol{n}, \boldsymbol{\beta}) \propto \pi(\boldsymbol{\beta} \mid \boldsymbol{\tau}^2)\, \pi(\boldsymbol{\tau}^2). \tag{10}$$

Since the full conditional distribution $\mathsf{p}(\boldsymbol{\beta} \mid \boldsymbol{n}, \boldsymbol{\tau}^2)$ in Equation (10) does not result in a closed form, a natural option to generate posterior samples is to resort to a Metropolis–Hastings algorithm with iteratively weighted least-squares (IWLS) proposals [18, 19]. In practice, a version of our model can be readily implemented with

the aid of existing statistical software. The S1 File includes examples with R code.

Results

We now apply our smooth multinomial logistic model to the GTC data. The masterplan of this section is as follows: first, we learn about the dynamics of the first-digit probability; second, we examine conformance of the first-digit probability to Benford's Law, and assess the homogeneity within the dataset over the observation period; third, we further examine evidence on the behavior of the second-digit probability. To streamline the comparisons with [8], below we partition the time horizon into two periods (S1: 1842–1960; S2: 1960–2010).

Dynamics of the First-Digit Probability

We present the dynamics of the probability $\boldsymbol{p}_t$ in Fig 3. The posterior mean of $p_{d,t}$ and the 95% credible band is compared to the corresponding probability from Benford's Law, along with the empirical distribution on each panel. The dynamics of posterior distributions of $p_{d,t}$'s show different patterns over the period between leading digits. As

Fig 3. Dynamics of the time-varying first-digit probability $\boldsymbol{p}_t$. Time-varying first-digit probability for digits 1 to 9 $(p_{1,t}, \ldots, p_{9,t})$ are presented from top left to bottom right. The chart further includes the posterior mean (solid line) and 95% credible bands (shaded areas) of $p_{d,t}$, the sample empirical distribution (point), and Benford's distribution (dashed line).

expected, we see that in the very early stage of the dataset, e.g. around the 1850s, the corresponding credible bands are much wider than those in the period of 1900s onward due to small sample sizes (see Fig 2). Among all the nine curves, the probability of leading digit one $p_{1,t}$ has a pronounced variation over the entire period. The posterior mean of $p_{1,t}$ rises to around 0.4 until the early 1910s, and then steadily drops for more than a century to around 0.2. This implies that the proportion of cyclones whose traveled distance start with digit one decreased approximately by half from around 1900s to recent years. On the contrary, the dynamics of the probability of leading digit three, i.e. $p_{3,t}$ moves upward the benchmark around the same period as the downward move of leading digit one, although the magnitude of the move is much smaller than that of digit one. The other seven curves move more tightly around the straight line of

Benford's Law, but digit two to digit seven are slightly upward and the others downward the benchmark. Given the variances of the digit probabilities, it is possible that these probabilities stay constant over the observation period S1.

Time-varying Conformance to Benford's Law

We now turn to time-varying conformance of the first-digit probability to Benford's Law. To assess overall conformance over nine digits with the First-Digit Rule in each year, we resort to the smooth SSD statistics from Equation 5. Fig 4 depicts the posterior mean and the 95% credible band of the smooth SSD. As with the first digit probability, the SSD also reflects uncertainty from different sample sizes and intrinsic variability of $p_{d,t}$'s.

The smooth SSD avoids overestimation of the misfit due to a discretization effect, whereas a naive empirical SSD as in [8] can be shown to be biased. As the numerical experiments in the S1 File illustrate, the empirical SSD can provide a biased and misleading snapshot of conformance to Benford's Law (see Fig 4, S1 File). For the GTC dataset, the empirical SSD (not reported) would be generally well above the smooth SSD curve from Fig 4, especially in the years where the number of cyclones was lower.

Fig 4. Dynamics of Sum of Squared Deviations (SSD). The chart gives the posterior mean of SSD (solid blue line) and 95% credible bands (shaded blue area) in each year. The time horizons suggested in the previous study is labeled for reference: Two long-term division (Period 1 and 2) and four short-term episodes (Episode A, B, C, and D).

The smooth SSD examines the heterogeneity within the dataset over time in terms of Benford's Law. Our results reject the hypothesis of homogeneity across the entire period of observation, as no horizontal line would fit the credible band of the smooth SSD. For the early decades prior to 1880s, the smooth SSD is susceptible to considerable variability due to small sample size, and hence it is difficult to tell either conformance or lack of conformance. However, ever since then, the posterior mean of the smooth SSD starts soon to increase gradually from the 1880s, reaches a peak value of 0.0184 in 1903, and then returns to a lower level around 1940, which constitute the first long-term cycle in the variation of the smooth SSD. Another substantive long-term deviation is currently in progress since the 1970s. The first peak occurs in 1989 with the posterior mean 0.00995 and then the mean falls slightly to 0.00757, ending up with the

highest value of 0.0153 in 2016. As shown in Fig 4, the second period has a large SSD value for the first period in magnitude, and hence these periods represent two different heterogeneity in the dataset.

To streamline comparisons, Fig 4 includes the sub-period division of [8]: Episode A and C show periods of decreasing misfit, which was claimed to be explained by technical advancements of collecting and coordinating data as a result of the introduction of telegraph lines and aircraft; Episode B, a sudden rise in a downward trend, was claimed to be possibly due by potential climate variation such as El Nino Southern Oscillation (ENSO); a small bump of misfit during Episode D was claimed to be possibly explained by a mix of effects of new technology and potential climate change.

Despite the conclusion of [8] that the GTC data tend to conform to Benford's Law from 1960 onward, our model actually finds a substantial deviation from Benford's Law over that period. Keeping in mind that technological improvements are cumulative, we find that the most recent heterogeneity levels actually tend to be higher than the ones from 1842 to 1890 (cf Fig 4). This finding seems to be contradiction with [8] (cf Fig 5 in their paper), which is possibly due to the above-mentioned bias issue faced by their approach.

Evidence from the Second-Digit Analysis

Fig 5. Dynamics of the time-varying second-digit probability $\boldsymbol{p}_t^{(2)}$. Time-varying second-digit probability for digits 0 to 9 ($p_{0,t}^{(2)}, \ldots, p_{9,t}^{(2)}$) are presented from top left to bottom right. The chart further includes the posterior mean (solid line) and 95% credible bands (shaded areas) of $p_{d,t}^{(2)}$, the sample empirical distribution (point), and Benford's distribution (dashed line).

We further examine the second-digit probability $\boldsymbol{p}_t^{(2)} = (p_{0,t}^{(2)}, \ldots, p_{9,t}^{(2)})$ in the GTC dataset. Benford's Second-Digit Rule is given by

$$\mathsf{p}_d^{(2)} = \mathsf{P}(D_2 = d) = \sum_{k=1}^{9} \log_{10}\left(1 + \frac{1}{10 \cdot k + d}\right), \quad d = 0, \ldots, 9, \tag{11}$$

where D_2 is the second significant digit of a random variable [20]. Fig 5 illustrates the dynamics of the second-digit probability $\boldsymbol{p}_t^{(2)}$. The dynamics of each $p_{d,t}^{(2)}$ yields either gradually increasing or decreasing linear trends over the entire period, but the variation

of each digit is mostly contained within the credible bands except for digit zero and digit four.

We also examine overall time-evolving conformance to Benford's Law between ten digits. The distribution of the second-digit smooth SSD is obtained from Benford's Second-Digit Rule in equation (11) and the posterior sample of the second-digit probability, and presented in Fig 6. The posterior mean of the second-digit SSD starts from high levels and dwindles until around 1950s, then gradually increasing up to recent years. The posterior mean of $\boldsymbol{p}_t^{(2)}$ gives consistent results to our finding in the first-digit analysis that the heterogeneity of the dataset may have been increasing recently.

Fig 6. Dynamics of the second-digit SSD. The chart presents the posterior mean of the second-digit SSD (solid blue line) and 95% credible bands (shaded blue area) in each year. The time horizons are labeled for reference as in the first-digit analysis

Closing Remarks

This paper devises a smooth Bayesian model based on penalized splines so to track time-varying conformance to Benford's law. We have explored the dynamics of the first- and second-digit probability to test the homogeneity of the GTC dataset by comparing the variation with Benford's Law. Our model enables us to track directly spans of years at which conformance to Benford's Law is lower, and therefore facilitates the statistical inference about the intrinsic distribution of the first or second digits by evaluating discrepancies from Benford's Law. Numerical studies in the S1 File show that our method avoids pitfalls faced by pointwise empirical approaches. With respect to our empirical findings versus those of [8]. There seems to be a consensus that the heterogeneity up to early 20th century could be mainly induced by the incomplete management of cyclone records and inevitable measurement errors. Technological developments in the 20th century have enable meteorologists to detect even tiny cyclones and to precisely locate the tracks of those cyclones, which results in the consistently increasing number of cyclones until the 1970s. Our results suggest that heterogeneity starts increasing again, even though the frequency of cyclones has been stable since the 1970s. While technological improvements are cumulative we find that the most recent heterogeneity levels actually tend to be higher than the ones from 1842

to 1890 (see Fig 4); this finding seems to contradict [8] (cf Fig 5 in their paper), possibly due to the above-mentioned bias issue.

While we have centered the paper on the tropical cyclone application, our Bayesian time-varying approach has the potential to be applied in other setups where the goal is on inferring the dynamics governing conformance to Benford's Law—including fraud analysis.

References

1. Newcomb S. Note on the frequency of use of the different digits in natural numbers. Am J Math. 1881;4(1):39–40.
2. Benford F. The law of anomalous numbers. Proc Am Phil Soc. 1938;78(4):551–572.
3. Hill TP. A statistical derivation of the significant-digit law. Statist Sci. 1995;10(4):354–363.
4. Miller SJ. Benford's law: theory & applications. Princeton NJ: Princeton University Press; 2015.
5. Tsagbey S, de Carvalho M, Page GL. All data are wrong, but some are useful? Advocating the need for data auditing. Am Statist. 2017;71:231–235.
6. Ley E. On the peculiar distribution of the U.S. stock indexes' digits. Am Statist. 1996;50(4):311–313.
7. Corazza M, Ellero A, Zorzi A. Checking financial markets via Benford's law: the S&P 500 case. In: Corazza M, Pizzi C, editors. Mathematical and Statistical Methods for Actuarial Sciences and Finance. Milano: Springer Milan; 2010. p. 93–102.
8. Joannes-Boyau R, Bodin T, Scheffers A, Sambridge M, May SM. Using Benford's law to investigate natural hazard dataset homogeneity. Scient Rep. 2015;5:12046.
9. R Development Core Team. R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing; 2016.

10. Emanuel K. Increasing destructiveness of tropical cyclones over the past 30 years. Nature. 2005;436:686.

11. Landsea CW. Hurricanes and global warming. Nature. 2005;438:E11.

12. Hijmans RJ. geosphere: spherical trigonometry; 2017. Available from: https://CRAN.R-project.org/package=geosphere.

13. Dobson AJ. An introduction to generalized linear models. 3rd ed. Chapman & Hall/CRC. Boca Raton: Chapman & Hall/CRC; 2008.

14. De Boor C. A practical guide to splines; rev. ed. Applied mathematical sciences. Berlin: Springer; 2001.

15. Fahrmeir L, Kneib T. Bayesian smoothing and regression for longitudinal, spatial and event history data. Oxford University Press; 2011.

16. Lang S, Brezger A. Bayesian P-splines. J Comput Graph Statist. 2004;13(1):183–212.

17. Brezger A, Steiner WJ. Monotonic regression based on Bayesian P-splines: an application to estimating price response functions from store-level scanner data. J Bus Econ Statist. 2008;26:90–104.

18. Gamerman D. Sampling from the posterior distribution in generalized linear mixed models. Statist Comput. 1997;7(1):57–68.

19. Brezger A, Lang S. Generalized structured additive regression based on Bayesian P-splines. Comput Statist Data Anal. 2006;50(4):967–991.

20. Diekmann A. Not the first digit! Using Benford's law to detect fraudulent scientific data. J Appl Statist. 2007;34(3):321–329.

Supporting Information

S1 File. Supplementary Materials.
(PDF)

Supplementary Materials

This supplement includes numerical experiments showcasing the performance of the methods and R code to implement the proposed approach along with some supporting reports on empirical results and Bayesian inferences.

Numerical Experiments

Simulation Configurations and Preliminary Experiments

To assess the performance of our method, we simulate data from

$$(n_{1,t}, \ldots, n_{9,t}) \sim \text{Mult}(30, p_{1,t}, \ldots, p_{9,t}), \quad t = 1, \ldots, 80, \tag{1}$$

where $(n_{1,t}, \ldots, n_{9,t})$ are the joint counts of leading digits with $\sum_{d=1}^{9} n_{d,t} = 30$ at time t, and where we assume that the time-varying first-digit probabilities are

$$p_{d,t} = \log_{10}\left[\frac{1 + 9 \cdot (d/9)^{\theta_t}}{1 + 9 \cdot \{(d-1)/9\}^{\theta_t}}\right], \quad \theta_t = 1 + 0.5 \cdot \sin\left(\frac{t}{10}\right), \quad d = 1, \ldots, 9. \tag{2}$$

Note that $\sum_{d=1}^{9} p_{d,t} = 1$, for every t. Fig 1 illustrates the dynamics over time of the true first-digit probability as in (2).

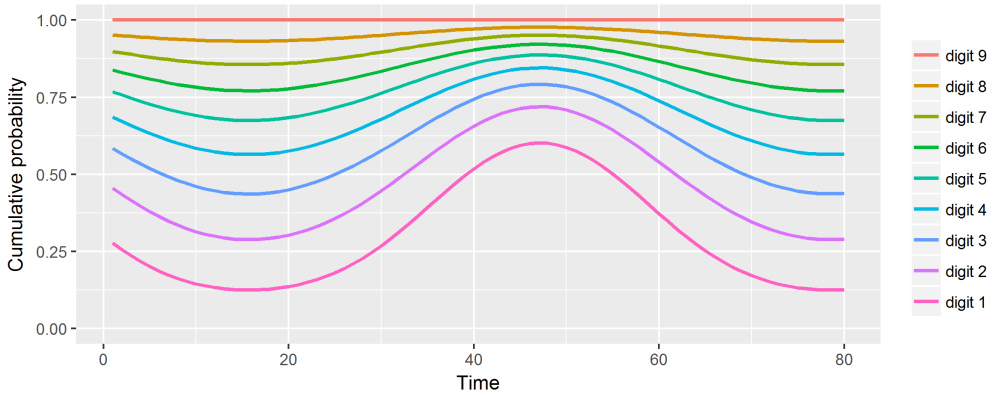


Fig 1. Dynamics of the first-digit cumulative probability. Each line represents the cumulative multinomial probability up to digit d, i.e. $\sum_{i=1}^{d} p_{i,t}$.

First, we concentrate on illustrating the method on a single-run experiment; Monte Carlo evidence is reported in the next section. We generate a random sample from (1),

and then apply our model to obtain posterior distribution of the first-digit probability. We run four chains of size 2,000 using Metropolis–Hastings algorithm with burning-in first 1,000 iteration and thinning 4. Fig 2 depicts the posterior mean of $p_{d,t}$, along with 95% credible bands and the true multinomial probabilities. As can be seen from Fig 2, the posterior mean of $p_{d,t}$ follows closely the true $p_{d,t}$ as defined in (2), and the credible bands tend to include the true $p_{d,t}$. Moreover, if the pooled dataset follows Benford's Law, we can make an inference on when the first-digit probability deviates from the first-digit rule by comparing the posterior distribution of the first digit probability mass with the horizontal line from Benford's Law. For now, the result should be regarded as

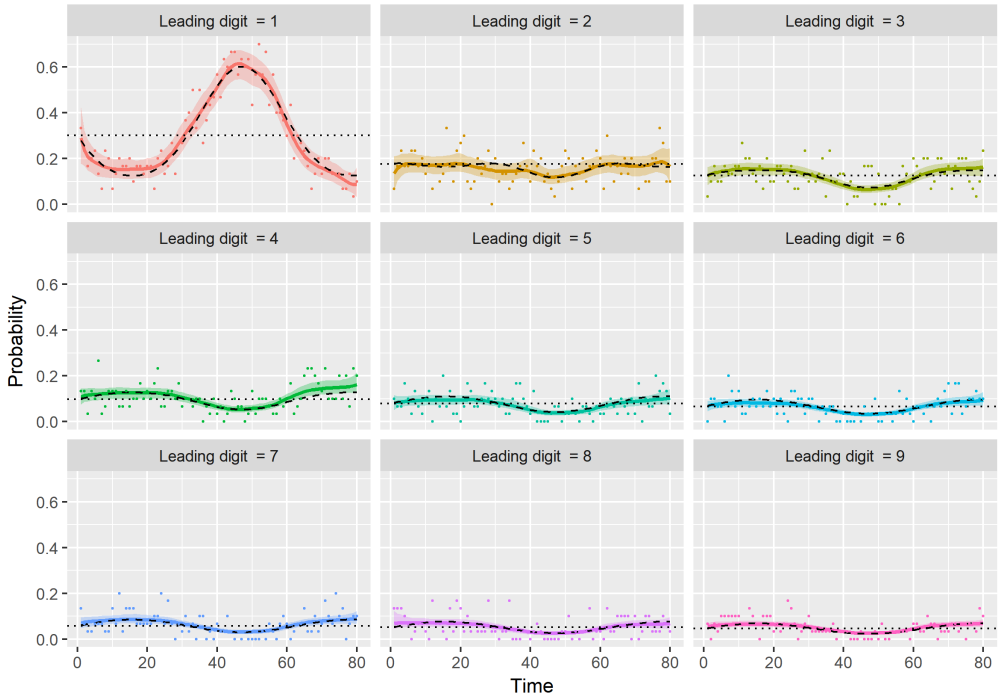


Fig 2. A single-run experiment with data simulated according to (2). On each panel, we represent the posterior mean of $p_{d,t}$ (solid line), the 95% credible bands (shaded area), empirical distribution (points), the true $p_{d,t}$ (dashed line), and the probability mass of Benford's Law (dotted line).

tentative, since Fig 2 summarizes the outcome of a single-run experiment. Next, we assess how robust these findings are over other runs of simulated data.

Monte Carlo Evidence. A Monte Carlo study was conducted by simulating $B = 500$ samples from the model in (1), using the same setting as in the previous section (that is, $N_t = 30$ and $p_{d,t}$ as in (2)). Fig 4 displays trajectories of the posterior means across 500

simulated datasets and their Monte Carlo mean. Our method successfully recovers the corresponding true first-digit probability, in spite of considerable variations of the multinomial probabilities over the period.

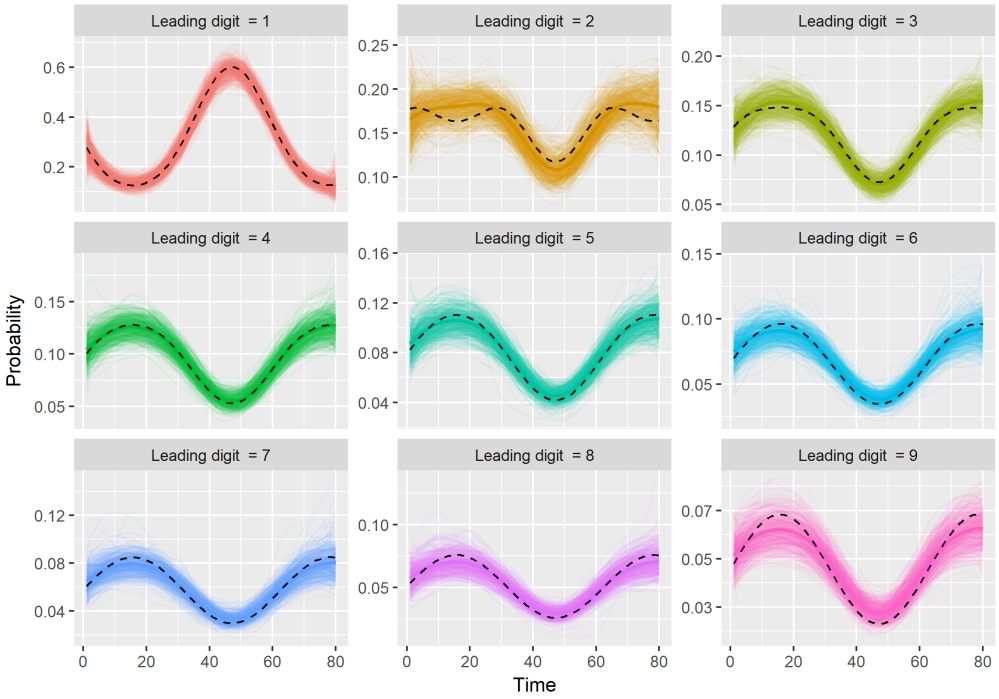


Fig 3. Trajectories resulting from fitting the model on simulated datasets and their Monte Carlo mean. On each panel, we present all the trajectories (translucent lines), the Monte Carlo mean (solid line), and the true $p_{d,t}$ (dashed line).

Discretization Effects. Fig 4 highlights that the empirical-based approach by [1] can suffer from bias.

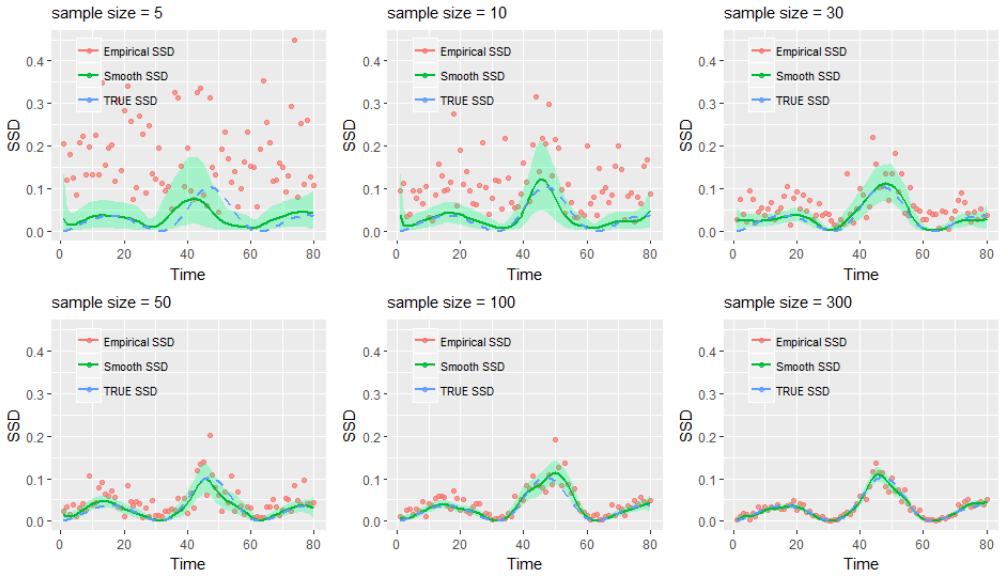


Fig 4. Sum of Squared Deviations (SSD) over six different sample sizes. On each panel, we present the true (blue line), smooth (green line), and empirical (red points) SSD.

R code

In this section, we present R code for implementing the time-varying model used in the Numerical Experiments. The interpretation of the results in the script is discussed in the previous section. In the code chunks below, we follow the 80 characters per line standard. Before running the code chunks, we start by installing the packages `splines2` and `R2jags` (if not installed). The `splines2` package yields B-splines basis functions and the `R2jags` package implements a Metropolis–Hastings algorithm by calling `JAGS` (Just Another Gibbs Sampler), a statistical software for Bayesian data analysis.

```
## Install required packages
packages <- c("R2jags", "splines2")
new <- packages[!(packages %in% installed.packages()[, "Package"])]
if (length(new)) install.packages(new)
## Load required packages
sapply(packages, require, character.only = TRUE)

##   R2jags splines2
##     TRUE     TRUE
```

Next, we define the true time-varying first-digit probability in (2) and then generate multinomial random vectors in (1) at time t using the `rmultinom` function. The seed (`set.seed`) is fixed below for reproducibility reasons.

```
## Define the true time-varying first-digit probability
t <- 1:80 # time span
d <- 1:9 # digits
N <- 30 # number of realizations at each time
theta <- 1 + 0.5 * sin(t / 10)
prob <- matrix(0, nrow = 80, ncol = 9)
for (i in t) {
  prob[i, ] <- log10(1 + 9 * (d / 9)^theta[i]) - log10(1 + 9 * ((d - 1) / 9)^theta[i])
}
## Generate a sample from the true time-varying probability
set.seed(789)
y <- matrix(0, nrow = 80, ncol = 9)
for (j in t) y[j, ] <- rmultinom(1, size = N, prob[j, ])
```

We then set the number of knots and compute B-spline predictors, and set the penalty matrix to use penalized splines.

```
## Setting up penalized splines
no.in.knots <- 15 # number of internal knots
in.knots <- quantile(t, # Generate equi-distant knots
                         probs = (1:no.in.knots) / (no.in.knots + 1), type = 1)
Bsp <- bSpline(t, knots = in.knots, degree = 3, intercept = TRUE)
Dd <- cbind(diag(length(in.knots) + 3), 0) - cbind(0, diag(length(in.knots) + 3))
Kmat <- t(Dd) %*% Dd # Penalty matrix
```

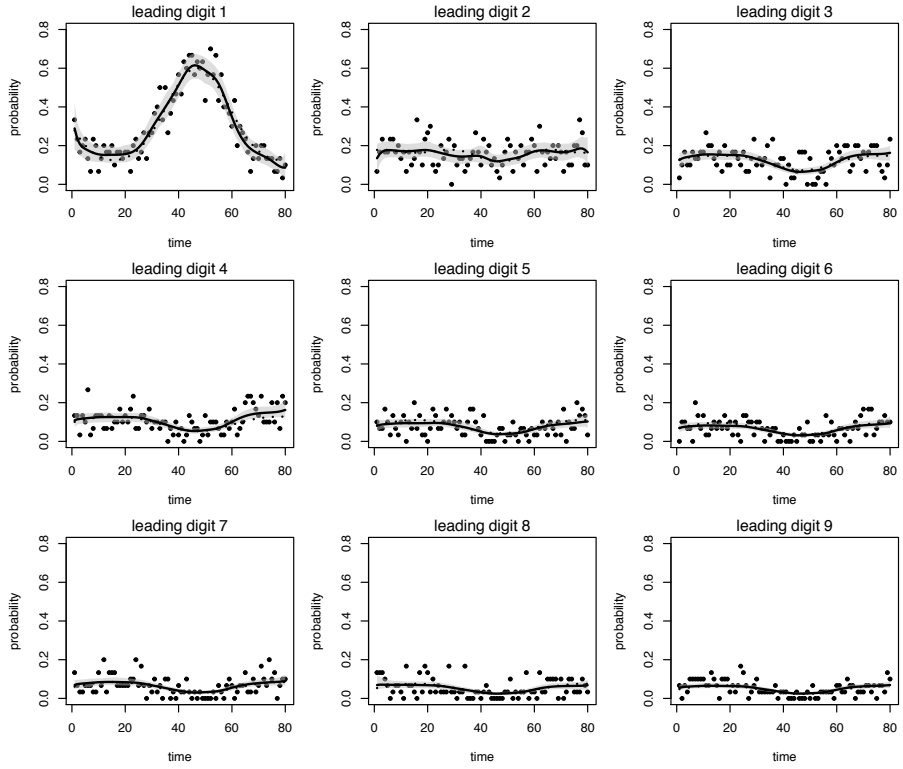
The following code chunks are used for calling and implementing our method in `JAGS`. In R, we can write the model in BUGS language and specify parameters, initial values, and data. The command `jags` connects inputs in R to `JAGS` and saves the simulations for easy access in R.

```
## Define objects for JAGS software
# JAGS model (BUGS language)
model <- function() {
  for (l in 1:8) {
    beta[1, l] ~ dnorm(0, 0.0001) # prior for beta1
    for (m in 2:(no.in.knots + 4)){ # random walk priors for beta's
      beta[m, l] <- beta[m - 1, l] + u[m - 1, l]
      u[m - 1, l] ~ dnorm(0, tau[l]) }
    tau[l] ~ dgamma(0.0001, 0.0001) } # prior for tau's
  for (i in 1:80) { # likelihood
    y[i, 1:9] ~ dmulti(pt[i, 1:9], N)
    for (j in 1:8) {
      eta[i, j] <- inprod(Bsp[i, ], beta[, j])
      eeta[i, j] <- exp(eta[i, j])
      pt[i, j] <- exp(eta[i, j]) / (1 + sumeeta[i]) }
    pt[i, 9] <- 1 / (1 + sumeeta[i])
    sumeeta[i] <- sum(eeta[i, ]) } }
# JAGS initial values for tau's
inits <- list( list(tau = rep(0.5, 8)), list(tau = rep(1, 8)),
               list(tau = rep(2, 8)), list(tau = rep(3, 8)))
# JAGS parmeters
parameters <- c("pt", "tau")
# JAGS data
data <- list("y", "N", "Bsp", "no.in.knots")
## Run JAGS in R
results <- jags(data, inits, parameters, model, n.chains = 4,
                                n.iter = 5000, n.thin = 10, n.burnin = 2500)
```

We now plot the resulting outcomes. Below, we present the empirical distribution as points, the posterior mean as a solid line, the credible bands as a polygon and the true multinomial probability as a dashed line.

```
## Extract MCMC samples
pt.array <- results[["BUGSoutput"]][["sims.list"]][["pt"]]
pt.mean <- apply(pt.array, c(2, 3), FUN = mean)
pt.ci <- apply(pt.array, c(2, 3), quantile, probs = c(0.025, 0.975))
```

```
## Plot the time-varying multinomial probabilities
par(mfrow = c(3, 3), mar = c(4, 4, 1, 0) + 0.5)
for (i in 1:9) {
  plot(t, y[, i] / N, type = "p", pch = 20, xlab = "time",
       ylab = "probability", ylim = c(0, 0.8))
  polygon(c(t, rev(t)), c(pt.ci[1, , i], rev(pt.ci[2, , i])),
          col = rgb(190, 190, 190, 127, maxColorValue=255) , border = FALSE)
  lines(t, prob[, i], lwd = 2, lty = 3)
  mtext(side = 3, text = bquote(leading ~ digit ~ .(i)), line = 0, cex = 0.8)
  lines(t, pt.mean[, i], lwd = 2)
}
```

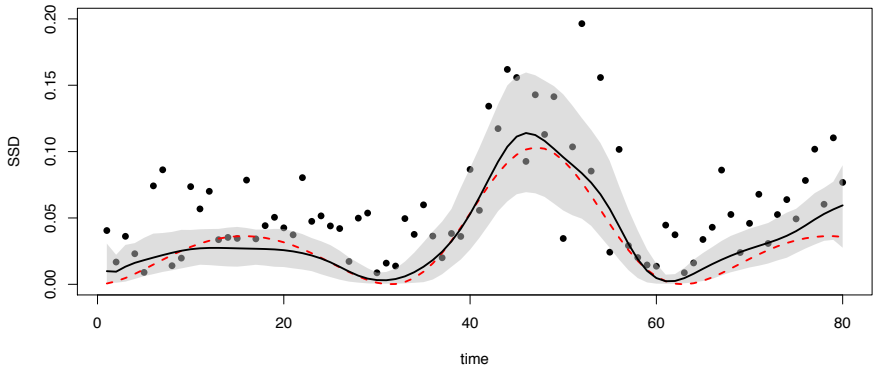


Finally, the code below can be used for comparing the sum of squared deviations (SSD)

among empirical distribution, the first-digit probability, and the true multinomial probabilities.

```
## Calculate SSD's
Ben.prob <- log10(1 + 1 / 1:9)
Ben.matrix <- matrix(Ben.prob, byrow = TRUE, nrow = 80, ncol = 9)
SSD.true <- rowSums((prob - Ben.matrix)^2) # true SSD
SSD.emp <- rowSums((y / N - Ben.matrix)^2) # empirical SSD
dev <- array(NA, dim(pt.array))
for (i in 1:9) {
 dev[, , i] <- pt.array[, , i] - matrix(Ben.prob[i], dim(pt.array)[1], dim(pt.array)[2])
}
SSD.dev <- apply(dev, c(1, 2), FUN = function(x) sum(x^2)) # smooth SSD
SSD.mean <- colMeans(SSD.dev)
SSD.ci <- apply(SSD.dev, 2, quantile, probs = c(0.025, 0.975))
par(mfrow = c(1, 1), mar = c(4, 4, 1, 0) + 0.5)
```

Below, we present the empirical SSD as points, the posterior mean of the smooth SSD as a solid line, the credible bands of the smooth SSD as a polygon and the true SSD as a dashed line.



Supporting Reports on Empirical Results and Bayesian Inferences

S Fig. Frequency of Cyclones and Relative Frequency of Traveled Distances

The chart shows the number of tropical cyclones since 1850 and the relative frequency of traveled distances in kilometers in each year.

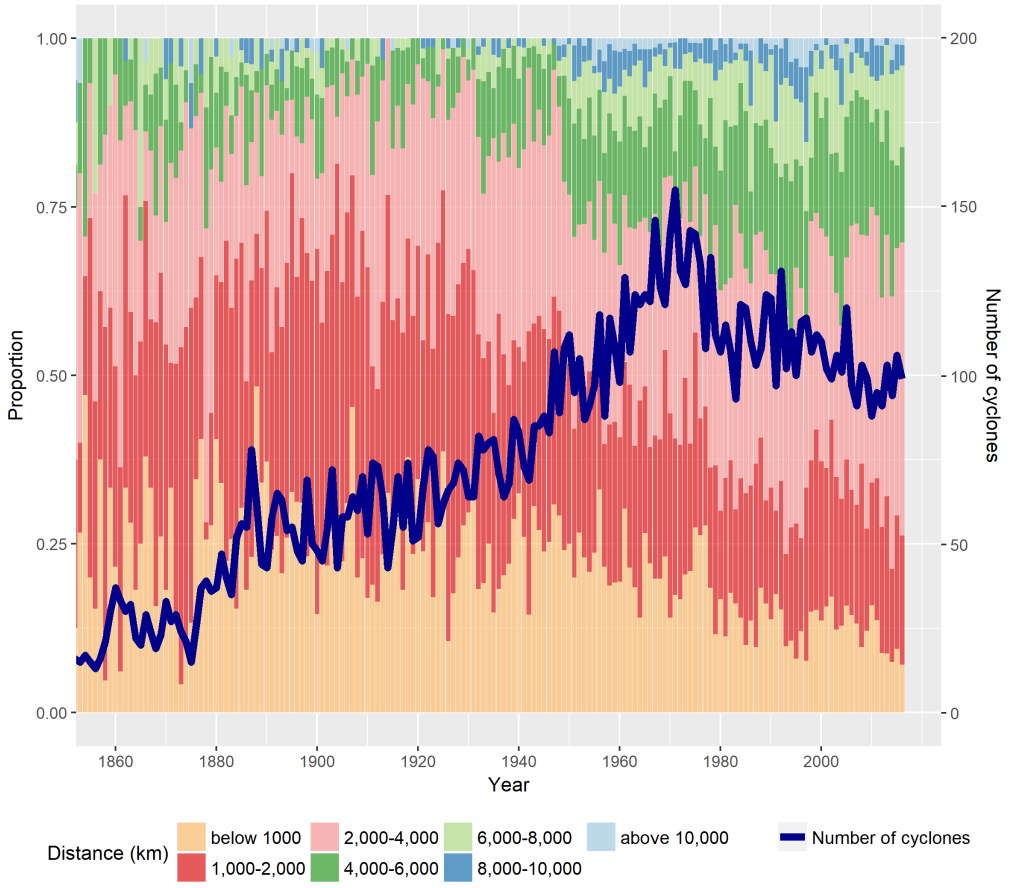


Fig 5. The frequency and relative frequency of traveled distances since 1850. The blue solid line depicts the number of tropical cyclones over time. In each year, the bars show the relative frequency of traveled distances in kilometers.

S1 Fig. Posterior Predictive Checks

This chart shows the posterior predictive distribution for the first-digit probability $\boldsymbol{p}_t$ from our model. As it can be observed, most observed proportions for each digit are covered by the respective 95% credible bands of the predictive distribution, thus suggesting that the model fits well the data.

Fig 6. Posterior predictive distribution and model fitting.
The posterior predictive distribution for each digit is presented over the period under analysis. The chart shows the posterior mean (solid line) and 95% predictive credible bands (shaded area), and the sample empirical distribution (point).

S1 Fig. Sensitivity Analysis

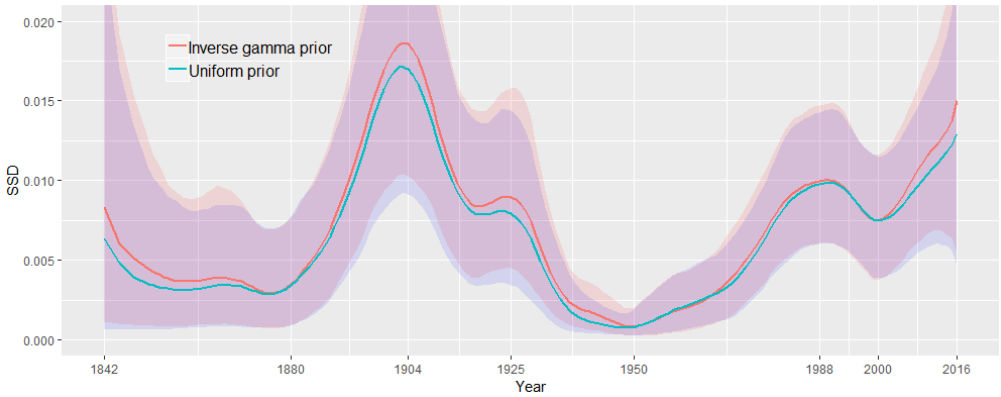


Fig 7. Sensitivity analysis with different priors. The chart compares the dynamics of SSD between two different priors for τ_d. The results from the inverse gamma prior (red) used in the paper are plotted against those from a uniform prior (blue).

References

1. Joannes-Boyau R, Bodin T, Scheffers A, Sambridge M, May SM. Using Benford's law to investigate natural hazard dataset homogeneity. Scient Rep. 2015;5:12046.